

Regression with Distance Matrices

Julian J. Faraway*

March 18, 2013

Abstract

Data types that lie in metric spaces but not in vector spaces are difficult to use within the usual regression setting, either as the response and/or a predictor. We represent the information in these variables using distance matrices. A low-dimensional representation of such distance matrices can be obtained using methods such as multidimensional scaling. Once these variables have been represented as scores, an internal model linking the predictors and the response can be developed. Here we use partial least squares. Scoring is the transformation from a new observation to a score while backscoring is a method to represent a score as an observation in the data space. Both methods are essential for prediction and explanation. We illustrate the methodology for shape data, unregistered curve data and correlation matrices using motion capture data from an experiment to study the motion of children with cleft lip.

Keywords: functional data analysis, mixed data, multidimensional scaling, shape, correlation matrix

1 Introduction

Regression methodology has been extended over the years to encompass a wider range of variable types. Categorical predictors have been represented using dummy variables and generalised linear modeling has allowed additional response types. More recently, functional data have been woven into the framework by representing functions using basis function expansions with the resulting vectors of coefficients now easily incorporated into the regression. Other types of data are more difficult to accommodate.

Data types such as unregistered curves, shapes, images, trees, covariance matrices etc. are difficult to integrate into the standard regression framework because such objects do not lie in a natural vector space. In some cases, linearisations can be achieved using basis function or tangent space representations but this is not simple. It is easier to define distances between such objects and hence define a metric space. In this paper, we envisage situations where the predictors and/or response are represented as distance matrices. We will demonstrate how regression can be performed with distance matrices. We show how predictions for new cases can be made and how we may interpret the relationship between the predictors and the response.

There is some prior work on the case where the predictors are expressed as a distance matrix. Cuadras and Arenas (1990) and subsequent papers present an approach to regression where the predictors are expressed as a distance matrix and principal coordinates analysis (PCO) is used to generate

*Department of Mathematical Sciences, University of Bath, BA2 7AY, United Kingdom, j.jf23@bath.ac.uk

scores. The response is then regressed on these scores. The advantage of the approach is that categorical predictors can be incorporated into the distance calculation in a way that differs from the usual dummy variables approach. Sometimes a better fit may be obtained from this approach. The drawback is that some of the explanatory value of regression coefficients in the standard approach is lost. Regression based on a distance matrix for the predictors can be seen in other approaches such as Gaussian Process Regression - at the heart of such methods there is a distance matrix of the predictors. See Rasmussen and Williams (2006) for more.

In other fields, the response is treated as a distance matrix. In Ecology, abundance matrices of species are sometimes converted to distance matrices because the standard Gaussian assumptions of MANOVA can not be justified. See McArdle and Anderson (2001) who show how the standard partitioning of sums of squares in MANOVA is still possible with only knowledge of the distance matrix of the response.

In Legendre et al. (1994) and Lichstein (2006), a modeling approach where both the response and predictors are represented as distance matrices is presented. However, the method unrolls the matrices, column by column, into vectors and then proceeds with regression which does not respect the geometry of the problem. This would not allow the extrapolation and interpolation necessary for the interpretation.

Although several papers have appeared treating either the response or the predictors as a distance matrix, the focus has been on hypothesis testing in the response case and prediction in the predictor case. Our focus in the paper is the explanatory value of regression in showing which aspects of the response depend on which aspects of the predictors. Certainly there are many “black box” methods that could be applied to some of the situations we shall describe but even if these achieve some measure of predictive performance, this is of little value if explanation is the main goal of the model.

In Section 2, we describe the method in general. In Section 3, we discuss an example with shape variables as both predictor and response. In Section 4, we demonstrate the modeling of an unregistered curve predictor and a shape response. Section 5 covers an example involving a correlation matrix as a response. The data in all three cases come from a study of children with cleft lip where the objective is to understand factors that affect the motion of the face. In Section 6, we end with some conclusions.

2 Methods

The proposed modeling process is illustrated in Figure 1. The predictors X will be expressed through a distance matrix D^X . Low dimensional coordinates S^X will be used represent D^X . A similar process will be used to form D^Y and S^Y from the response Y . A linear model will be used to link S^X and S^Y . By itself, this linear model can only measure the strength of the relationship between X and Y . To make the method useful in practice, we describe how to map between the spaces of X and S^X and similarly for Y and S^Y . This enables interpretation and prediction on the scales of X and Y .



Figure 1: Predictors X form distance matrix D^X from which scores S^X are calculated using MDS. Similarly for the response Y . The scores are related using PLS.

2.1 Distance matrices and scores

Consider a metric space (M_X, d_x) where distance d is a function such that $d_x : M_X \times M_X \rightarrow \mathbb{R}$. The choice of distance is critical to the outcome and the user must choose carefully according to the nature of X . Finding a suitable distance may not be easy, particularly on difficult manifolds. Given a sample of data x_1, \dots, x_n , we compute an $n \times n$ distance matrix D^x where $D_{ij}^x = d(x_i, x_j)$. Similarly, we have another, and possibly quite different, metric space for the response, (M_Y, d_y) and associated distance matrix D^y .

Classical multidimensional scaling(cMDS) forms a matrix B from distance matrix D such that:

$$B_{ij} = -(D_{ij}^2 - D_{i.}^2 - D_{.j}^2 + D_{..}^2)/2$$

where the dots in the subscripts indicate that means are taken over the index. We form the eigen-decomposition: $B = SS^T$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \lambda_n$. The columns of matrix S contain the principal coordinates or scores. We perform this decomposition on both D^x and D^y to obtain S^x and S^y respectively. There is some choice regarding the dimension of S^x and S^y . Since there will be some later dimension reduction in the modeling, we should be inclusive in this selection. Nevertheless, the computational burden can be reduced by eliminating dimensions of the scores which correspond to relatively small eigenvalues. There are many other ways of forming low-dimensional coordinate representations of distance matrices, for example ISOMAP (Tenenbaum et al. (2000)), that might, in principle, be used here but the worked examples use cMDS.

2.2 Internal Linear Model

We need to model the relationship between S^x and S^y . There are several suitable choices. We choose the method of Partial Least Squares(PLS) using an implementation obased on Mevik and Wehrens (2007) using the SIMPLS method of de Jong (1993). PLS finds the linear combinations of S_X that are most strongly related to S_Y . PLS has well-established variable selection methods based on cross-validation which determine the size of the model that will give the best predictive behaviour.

Various alternatives might be considered for modeling the relationship between the scores. Multivariate Multiple Regression could be used, but PLS is generally able to obtain a superior fit to the data. PLS regression is harder to interpret because of the linear combinations but since our explanations will only be interpretable in the data space rather than the score space, this loss is not a concern. Principal Components Regression could also be used here but PLS is usually able to better fit the relationship between the predictors and the response. Canonical Correlation analysis could also be used but this is a regression rather than a correlation problem so that an asymmetric method like PLS is preferable. Nevertheless, there are many other methods that could be substituted here and we claim no general superiority for PLS.

2.3 Scoring and Backscoring

Suppose a predicted response is required for a new value of x . Following the modeling process shown in Figure 1, we need to convert this x into a score (which we call *scoring*). The internal linear model will then be used to predict a response score. This score will then need to be converted back to the space of the response (which we call *backscoring*). Scoring and backscoring are essential for the interpretation of the model in the spaces of the observed variables. This will usually require scoring and backscoring for both the predictors and the response.

Consider a new data point x_{new} which we must map into the score space. Appending x_{new} to the observed x 's and recomputing the distance matrix will change all the previously computed scores which is not viable. We must maintain the existing scores while locating the new score that reflects the distances of x_{new} to the observed x 's. Gower (1968) describes one way this may be achieved:

$$s_{new}^x = \Lambda^{-1} S^T (d^2(\mathbf{x}, \bar{x}) - d^2(\mathbf{x}, x_{new}))/2 = \phi(x_{new}) \quad (1)$$

where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ and \bar{x} is the centroid. However, $d^2(\mathbf{x}, \bar{x}) = -\text{diag}(B)$ so explicit computation of the centroid is not yet needed. This computation requires only the distances $d(\mathbf{x}, x_{new})$ of the new point to the original points. In practice, we retain only the first few coordinate directions so appropriately reduced versions of Λ and S would be used, and s_{new} represents a projection onto the reduced score space. Other methods of scoring based on various motivations have been proposed. As discussed in Faraway (2012), several of these methods are equivalent to Gower's method.

Backscoring is more problematic. The internal linear model will produce s_{new}^y which we want to map to the space of the observed response. Each s_{new}^y will correspond to a set of values in M_y which have this score. Within this set, a natural choice is the one closest to the centroid, that is:

$$y_{new} = \arg \min_{s_{new}^y = \phi(y)} d(y, \bar{y}) \quad (2)$$

This does require the computation of a centroid in the response space for which we may use the Fréchet mean:

$$\text{argmin}_{y \in M_y} \sum_{i=1}^n d(y, y_i)$$

Finding a solution to (2) is difficult when the data do not lie in a vector space because most optimisation methods need this property. In a Euclidean space, the MDS is effectively equivalent to a principal components analysis and the solution is explicit. But there would be little incentive to use our method for Euclidean spaces as more direct methods apply. The method proposed here is only valuable for the more difficult types of data for which the optimisation is problematic. The constraint that $y \in M_y$ further complicates the search for a solution which needs to be customised to the M_y . We will describe such methods in the three examples to follow. Further discussion and examples of scoring and backscoring may be found in Faraway (2012).

2.4 Adding Predictors

Suppose we wish to add a predictor Z to the set of predictors lying in M_x . To maintain the spirit of the approach explained above, we might seek to modify the distance measure d_x to $d_{x,z}$ to accomodate the additional predictors. However, when X and Z are quite different in nature, the construction of an appropriate distance would be difficult. Alternatively, we may have a distance measure on Z and be able to obtain a distance matrix which might be combined with D_x using

$$D_{combined} = \sqrt{(D_x^2 + D_Z^2)}$$

However, there is the problem the relative weighting of X and Z in the combined distance measure which an ordinary regression would be represented by a parameter. This is difficult to incorporate here so a different approach is recommended.

Consider the case where Z comprises the usual quantitative and/or categorical predictors. We can use the standard regression methods to produce a model matrix S^Z which can be adjoined to S^X and

the PLS modeling can then proceed. A more difficult case arises where Z , like X also lies in a non-vector space and we must resort to a distance matrix D^Z . An MDS then produces a matrix of scores S^Z which can then be adjoined to S^X . PLS modelers commonly use dummy matrix representations of categorical variables in their inputs and outputs so our proposed approach fits well with this practice. Difficulties may arise in the interpretation of the fitted model since we must distinguish the effect of X from the effect of Z . But these problems arise also in standard regression situations so there is no additional difficulty in this setting.

2.5 Inference and Diagnostics

Inference and model checking can be made using the internal linear model. For methods such as PLS, there are well-established methods of making this inference and we have nothing to add to this. However, we may wish to assess and check the fit in the scale of the response. We have residuals given by $d_y(\hat{y}_i, y_i)$ which may be summed to form a residual sum of squares (RSS). Suppose we wish to compare a larger model with smaller, nested model then we can compute an F-statistic of the form:

$$F = (RSS_{small} - RSS_{large}) / RSS_{large}$$

The significance of this F-statistic can be explored numerically using permutation tests which we demonstrate in the examples to come. We can also compute an R^2 -like statistic of the form $1 - RSS/TSS$ where TSS is the total sum of squares derived from a null model of a constant predictor. The residuals may also be used to make diagnostics analogous to those used in common regression.

3 Shapes on Shapes

We consider shape as the information in a configuration of landmarks that is invariant to rotation, translation and size transformations. The data for the example comes from a subset of a study described in Trotman et al. (2010). 38 markers are placed in designated positions on the faces of 36 normal children as shown in Figure 2 and motions in 3D are recorded for four second segments at 60Hz. Among other actions, the children are asked to smile (starting from and returning to rest). For the purposes of this example, we consider only the initial pose, defined by the average position during the first 10 frames of motion and the maximal pose, defined by the frame of motion which is furthest from the initial pose in terms of Riemann distance. We consider the problem of relating the maximal pose described by a 38×3 shape matrix to the initial pose also described by a 38×3 shape matrix. We are interested in whether the rest pose of the face has any effect on the type of smile subsequently made.

We construct the distance matrices using the Riemann distance and then use MDS to construct the scores as described in the Section 2.1. These scores are approximately equivalent to the tangent space representation, provided the shapes are relatively concentrated. The distance and tangent space approximation are described in Dryden and Mardia (1998) with numerical implementation being provided by an R package called *shapes* from Dryden (2009). Faces do not change shape that much during motion so the shapes are concentrated in this instance. This makes the necessary mappings between the score and data spaces very much easier to compute because linear mappings can be used. If the shapes were more dispersed or we chose another distance function, as explored in Faraway (2012), the backscoring becomes more difficult.

We consider the percentage of the variation explained by choosing the dimension of the score space. For both the predictor and the response, this variation drops off relatively slowly with only

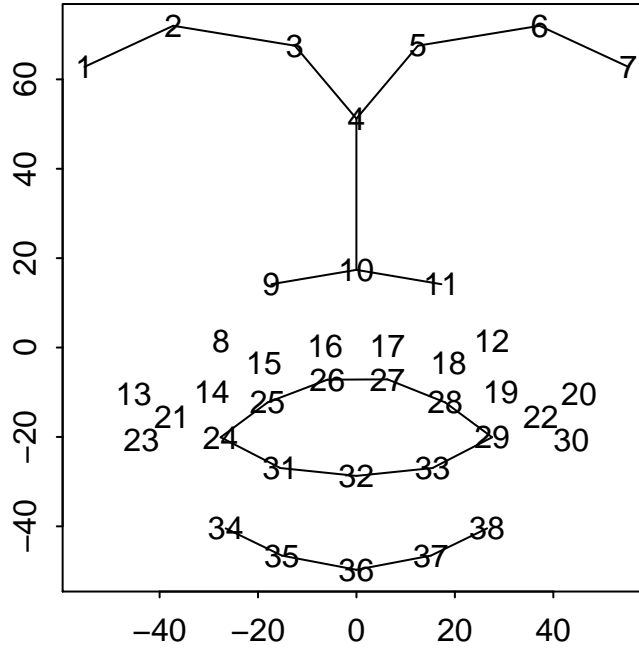


Figure 2: Configuration of 38 markers on the average face. Lines depicting the eyebrows, nose, mouth and lower jaw have been drawn.

20-25% of the variation explained by the first component. We choose 10 dimensions in both spaces which ensures that no unchosen component explains more than 2-3% of the variation.

In this application, our interest lies in explanation rather than prediction. For brevity in the exposition, we will focus on the first component of the response fitted by the PLS model. At least the second component of the response is of interest and could be pursued. Once we focus on the first dimension of the response, we find that, using leave-out-one crossvalidation, only one linear combination of the predictor scores is sufficient explanation. There is a correlation of 0.78 between these predictor and response scores indicating a strongly significant relationship. However, given the selection effect of choosing the maximally correlated linear combinations, we must be cautious about assuming this implies a strongly significant relationship between the shape predictors and response.

The scores can be mapped back to shapes using the appropriate linear combinations in the tangent space. These are depicted in Figure 3. In the top panel on the left, each arrow starts from the mean initial pose (corresponding to the markers in Figure 2) and ends at the location of that marker when perturbed up by two standard deviations in the direction of the first component. We can see this perturbation from the mean face corresponds to a pose where the lips are drawn in towards a kiss-like pose. On the top right, the contrasting perturbation down where we see the lips are drawn out into a thin grin. Note that motion is not being depicted in these plots, rather the variation about the mean pose and that the distinction between perturbing up and down is arbitrary.

These input scores can be related to predicted output scores which can again be represented in the shape space of the response as seen in the lower two plots of Figure 3. Note that the principal coordinate decomposition only reliably defines axes of variation so “up” and “down” are arbitrary. Even so, once the two PCOs have been made, we must take care to be consistent in the interpretation of the direction hence the left panel above does correspond to the left panel below. In the lower left panel, we see that this direction of variation corresponds to a mouth open smile where the teeth would likely be visible while on the lower right, we see a closed mouth smile where the lips are kept together

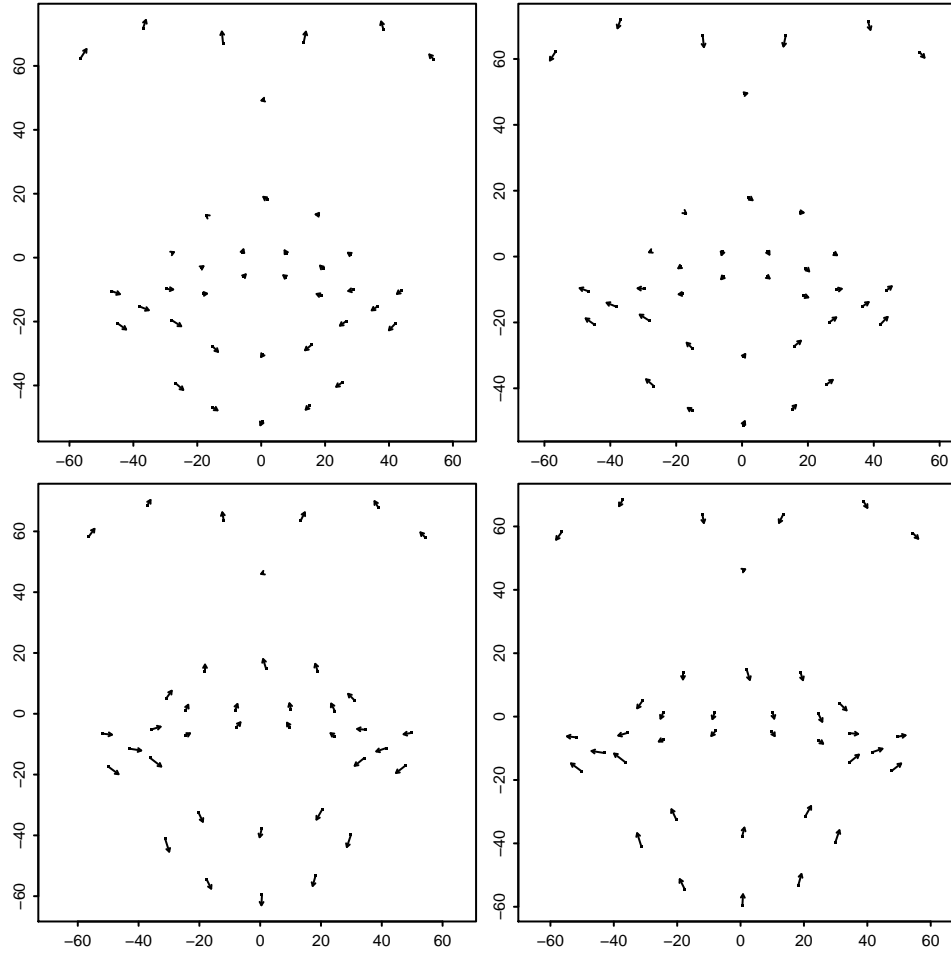


Figure 3: Arrows for each facial marker start from the mean pose and end perturbed by two SDs. The upper two plots show perturbations on the initial pose while the two lower plots show perturbations on the maximal pose. The plots on the left show perturbations up from the mean — we see a pursed lip initial pose leads to a mouth-open smile. The plots on the right show perturbations down from the mean — we a thin grin initial pose leads to a mouth-closed smile.

while the corners of the mouth move outward and upward. Thus the analysis shows that an initial pose closer to a kiss leads to a mouth open smile while a thin grin leads to a broader grin. Hence we see that the initial pose indicates some anticipation of the type of smile to come.

We can compute a test of significance for the predictor in this regression. The test statistic has a value of 0.179. Using a permutation test (where the response scores are permuted), we determine that the p-value is very small (none of the 1000 permuted test-statistics exceeded the observed value).

4 Shapes on Curves

There is a growing literature on functional data analysis. Substantial work has been done to integrate one-dimensional functions both as predictors and responses in a regression problem. The usual approach is approximate the functions as linear combinations of basis functions. The coefficients of these linear combinations are vectors which can be readily integrated into the standard regression framework. See Ramsay and Silverman (2005) for an overview. A common preliminary step in a functional data analysis is registration where phase variation in the observed curves is removed. In the analysis to follow, we retain both the phase and amplitude variation in the observed curves as both forms of variation may be of interest.

Consider the same set of data as in the previous example except in this case we compute the Riemann distance of each frame in the motion to the initial shape. This produces a curve which will be used as the predictor in this example and are shown as the grey curves in the first panel of Figure 4. The curve represents both the timing (phase variation) and the magnitude (amplitude variation) for the motion. For now, we do not include the initial pose as a predictor. Instead, we standardise the response as

$$s_{std} = s_{max} - s_{initial} + s_{mean}$$

where the shapes s (respectively, standardized, maximal pose, initial pose, mean (over subjects) pose) are represented in a tangent space where addition and subtraction are meaningful. The purpose of this transformation is to remove the variation in the rest poses of the faces by estimating the maximal pose that would be achieved if started from the mean initial face.

We use the L_1 distance (area) between curves defined by:

$$d(f_1, f_2) = \int |f_1 - f_2| dx$$

We could scale this on the horizontal (time scale) and/or vertical (magnitude scale) but these scale factors will not affect the analysis. We apply multidimensional scaling to this distance matrix and extract the first two components as representative of the motion. This L_1 distance is simpler than that described by Faraway (2012) where a dynamic time warping distance was used.

As before, we use the tangent space coordinates for the shape space to represent the response. Again we choose 10 components. We then apply PLS as described earlier and find that two components are helpful in describing the relationship between the predictors and response.

We now describe how we obtain a mapping between the score space of X and the original function space. To solve the optimization problem given by (2), we must find some way to explore the manifold of functions that represent possible motion patterns. Although we can see from the grey curves in the first panel of Figure 4 that these functions have some common form, the manifold is difficult to describe explicitly. We could parameterize but we prefer to seek a nonparametric solution with fewer assumptions.

We develop a method of forming weighted combinations of two or three functions. By combining only a few functions, we hope to retain those properties of the functions which make them members of the implicit manifold. Averaging over many functions runs the risk of producing smooth functions which lie outside the space - for example, consider the pointwise average of all functions.

We define a family of transformations of a curve that may be viewed as defining tangents to the space at that point. Consider mappings of the form: $f(t) \rightarrow f^p(t, \mathbf{p})$:

$$f^p(t, \mathbf{p}) = \begin{cases} p_3 f(t p_1 / p_2) & t \leq p_2 \\ p_3 f((1 - p_1)t / (1 - p_2) + (p_1 - p_2) / (1 - p_2)) & t > p_2 \end{cases}$$

where $p_1, p_2 \in [0, 1]$ and $p_3 > 0$. In words, pick a point p_1 in the interior and move it to p_2 , linearly rescaling either side. Scale by p_3 .

Now to form the weighted combination of two functions f_1 and f_2 , choose \mathbf{p}^1 so that the distance $d(f_1^p, f_2)$ is minimized and vice versa, reversing the role of f_1 and f_2 to obtain \mathbf{p}^2 . We form the γ -weighted average by weighting the scaling parameters p_2 and p_3 and averaging the resulting two functions:

$$\gamma f_1 + (1 - \gamma) f_2 \equiv \{f_1^p(p_1^1, (1 - \gamma)p_2^1 + \gamma p_1^1, (1 - \gamma)p_3^1 + \gamma) + f_2^p(p_1^2, \gamma p_2^2 + (1 - \gamma)p_1^2, \gamma p_3^2 + (1 - \gamma))\} / 2$$

As γ varies from 1 to 0, it forms an approximate geodesic between f_1 and f_2 . We can form the weighted average of three functions by forming a weighted combination of the first two functions and then combining that with the third function with the appropriate weights. The lack of invariance to the ordering of the functions is inelegant but makes little difference in practice.

The centroid \tilde{f} is now computed using the following algorithm:

1. For all triples of functions, find the weighted combination that has score $\mathbf{0}$ using Gower's scoring method. There may not be a solution for all triples.
2. For each feasible solution f^* , compute $\sum_i d(f^*, f_i)$. Average over the k best as the centroid. We chose $k = 10$ to achieve a smoother results without losing features.

For larger datasets, it will not be possible to compute all triples. A random sample can be used or we can restrict attention to functions whose score is close to $\mathbf{0}$. Some additional flexibility is possible by allowing the weights γ to vary slightly outside of $[0, 1]$.

To obtain the best function corresponding to a target score of s^* :

1. For all triples of functions, find the weighted combination that has score s^* . Solutions will be found only for some triples.
2. For each feasible solution f^* , compute $d(f^*, \tilde{f})$. Average over the k best to obtain estimated function.

It is possible that for s^* far from $\mathbf{0}$ that no solutions may found. It is in such cases that allowing the weights γ to vary slightly outside of $[0, 1]$ may be helpful although it simply may not be possible estimate a function with too large a score. An alternative to averaging over the k best estimates is to plot the estimates as this provides some notion of the variation among estimates with score s^* .

We see in Figure 4, that a shorter, smaller smile is associated with the lips-together smile while the longer, larger smile is associated with the mouth-open smile.

The second component of variation is seen in Figure 5. In this case, more of the response variation is visible from the side of the face rather than the front of the face. We see that a smaller, later smile

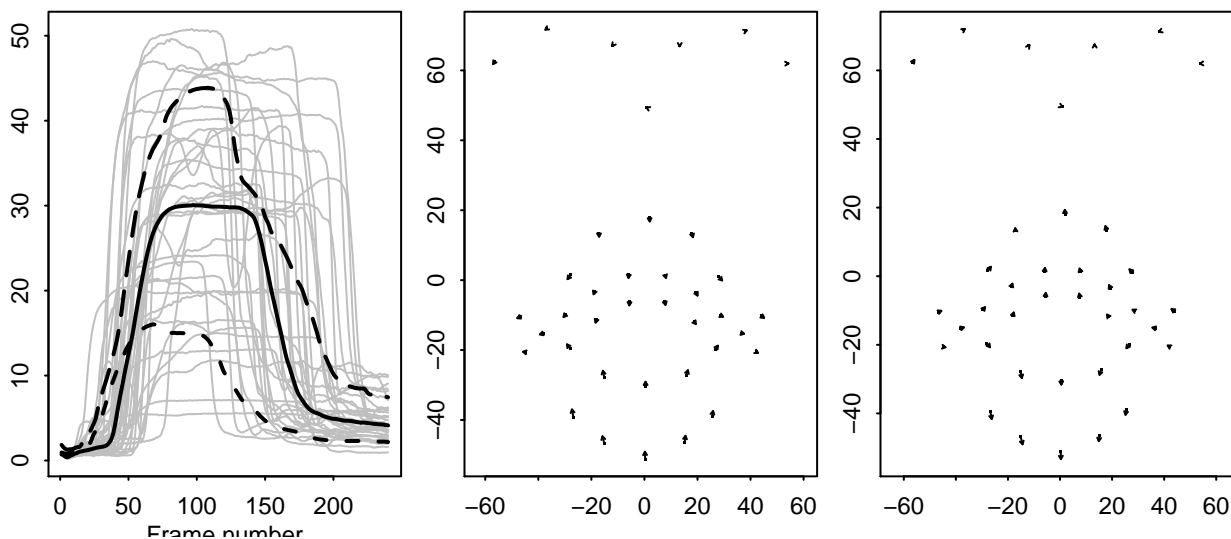


Figure 4: First plot shows the first component of the curve predictor variation. The data curves are shown in grey, the mean as a thick black curve and the dashed lines show two sd variations around this mean in the axis of the first component. The second plot shows the variation from the mean shape corresponding to the lower of the two dashed curves in the first plot. The third plot shows the variation around the mean shape corresponding to the higher of the two dashed curves in the first plot.

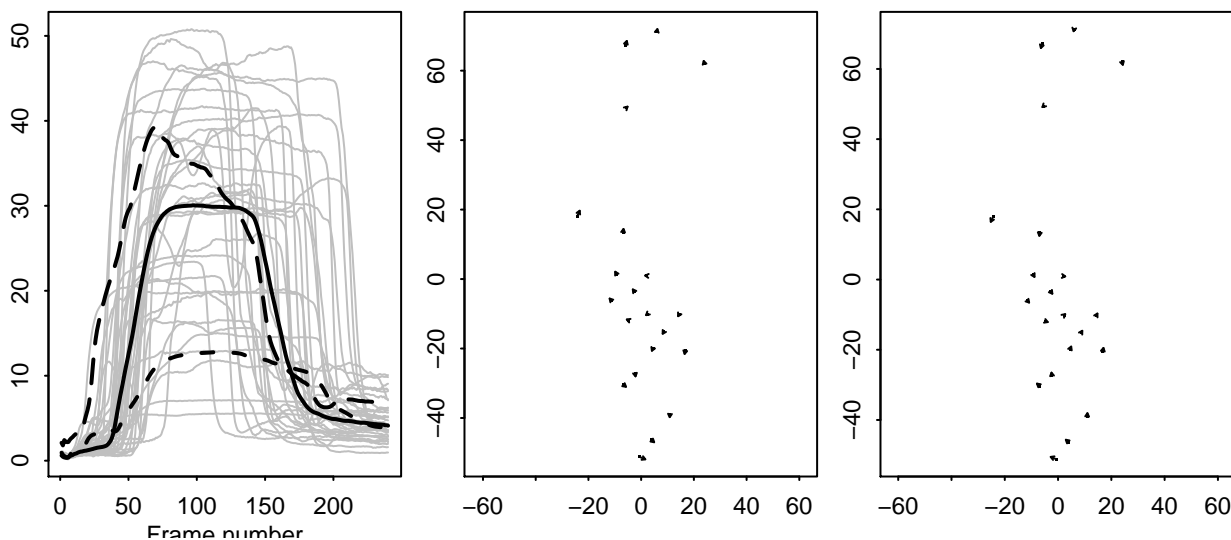


Figure 5: Same as Figure 4 except for the second component of variation. The side rather than frontal view is shown.

is associated with lips out, chin in smile while a larger, sooner smile is associated with lips back, chin out smile.

We can compute a test of significance for the predictor in this regression. The test statistic has a value of 0.725. Using a permutation test, we determine that the p-value is very small (none of the 1000 permuted test-statistics exceeded the observed value). Although the effects are statistically significant, the size of the effect is smaller than that observed in the first example where initial pose was used as a predictor. We now explore the relative importance of these potential predictors.

We use the initial pose, as described in the first example, the time-course function and the centroid size of the initial face, combining these predictors as described in Section 2.4. We compute the value of R^2 for all subsets of size two or more as seen in Table 1. We see that the initial pose is the most important predictor of the maximal pose but we may question if the time and size variables are needed in addition to the initial pose.

Model	R^2
initial + time + size	55.4
time + size	10.9
initial + time	54.6
initial + size	52.8

Table 1: R^2 for models predicted maximal pose.

We test for the significance of the size variable using the F-statistic and assessing the significance using a permutation test that permutes only the size variable for each resample. We obtain a p-value of 0.58 from 1000 replications and conclude that the size does not have a statistically significant effect relative to this model. Further, we test the significance of both the time and size variables using the same technique. Here we find a p-value of 0.78, indicating that we may predict the maximal pose with the initial pose alone.

5 Correlation Matrices

Correlation (or covariance) matrices can arise as variables in a regression analysis. Examples include radio communication as in Herdin et al. (2005), brain imaging as in Dryden et al. (2009) and longitudinal data modeling as in Daniels and Pourahmadi (2002). The space of positive semi-definite symmetric matrices is not Euclidean which poses difficulties for analysis. For example, the simple entrywise average of two correlation matrices may not itself be a correlation matrix. Although parametric modeling with Wishart distributions may be possible, this is difficult and we present a simpler, non-parametric approach here.

Given a sample of $m \times m$ correlation matrices C_1, \dots, C_n , we can compute a distance matrix D_{ij} where $i, j = 1, \dots, n$. Various distances for correlation matrices have been proposed, for example, see Herdin et al. (2005) but we shall use the Frobenius distance in the application to follow. Whether the correlation matrices appear as responses or predictors, we may apply the principal coordinates analysis as described above. However, useful interpretation requires that we be able to backscore to the correlation space once the internal modelling has occurred. We must first compute the Fréchet mean:

$$\bar{C} = \arg \min_C \sum_{i=1}^n d(C_i, C)$$

The optimisation is not straightforward because the space is non-Euclidean so we take a different approach. We define the weighted combination of correlation matrices as

$$C_\gamma = \arg \min_C d\left(\sum_{i=1}^n \gamma_i C_i, C\right)$$

which may be computed for the Frobenius norm using the algorithm due to Higham (2002). Note that $\sum_{i=1}^n \gamma_i C_i$ may be a correlation matrix in which case the solution is C . If it is not, we find the closest (in Frobenius norm) correlation matrix C . The implementation of the backscoring method requires a method to search over weighted combinations of C_i . Given a new score s we may use a Lagrange-like method to find:

$$C(s) = \arg \min_{C_\gamma} \{d(\bar{C}, C_\gamma) + \delta \|\phi(C_\gamma) - s\|\}$$

where δ is chosen in a balanced manner to allow minimisation of the first argument while enforcing the constraint of the second. Dryden et al. (2009) discuss the use of other metrics which avoid the projection used here to work with the Frobenius norm but would be more difficult to implement for this purpose.

We illustrate regression modelling using correlation matrices using the same data discussed in the earlier examples. Muscle coordination during facial motion is important in expression. In the study of patients with cleft lip and palate repairs, researchers would like to discover factors that affect this coordination. The markers on the face move in three dimensions, but in most cases, the motion is close to linear. We preprocess the data, first using generalised procrustes analysis to remove whole head motion and ordinary procrustes analysis to rotate the head onto a symmetrised face-forward configuration of the markers. The purpose of this step is to ensure that all 36 motions are face-forward without any whole head motion.

For each of the 38 markers on the face, we have a 240x3 matrix describing the trajectory for each subject. We perform a location standardisation so that all these trajectories start from the origin. We then concatenate these location shifted matrices into a (240x36)x3 matrix and perform a principal components analysis. The first principal component identifies the major axis of motion for that marker. We compute the scores corresponding to this component which show the timing and magnitude of motion along that axis. Hence for each subject, we may produce a 240x38 matrix of scores from which we may compute a correlation matrix. The correlations between the markers represent the extent to which the motions of the markers are correlated. We considered computing correlations directly on the 240x(38x3) observed motion matrices but the cross-correlation between the three coordinate directions would have been difficult to interpret.

We considered two covariates as possible predictors of these correlation matrices: the size of the face and speed with which the motion was performed. We use the centroid size which is related to the age of the child. We computed the Riemann distance of each frame of motion from the initial frame of motion and use the maximum rate of change in this distance (slightly smoothed) as an estimate of speed. We regressed the size and speed on the first two coordinates of the PCO and found that only the speed was a significant predictor of the first principal coordinate.

Now consider an internal model regressing the first principal component on the speed alone. Contrast the predicted response for the minimum observed speed with the maximum observed speed. The predicted scores can be used to construct predicted correlation matrices. In Figure 6, we can see the estimated mean correlation matrix which reveal many of the observed correlations are large but with some exceptions. The contrast in the correlations corresponding to the maximum and minimum observed speeds show that this predictor has little effect for many of the markers but that for markers

16, 17, 26 and 27 which are situated on the upper lip just below the nose (see Figure 2) and for markers 32,34 and 38 which are found on the jaw, there is a more substantial effect due to speed. At lower speeds, there is less coordination with other markers for these particular markers while at the higher speeds, these markers move more in conjunction with the other markers.

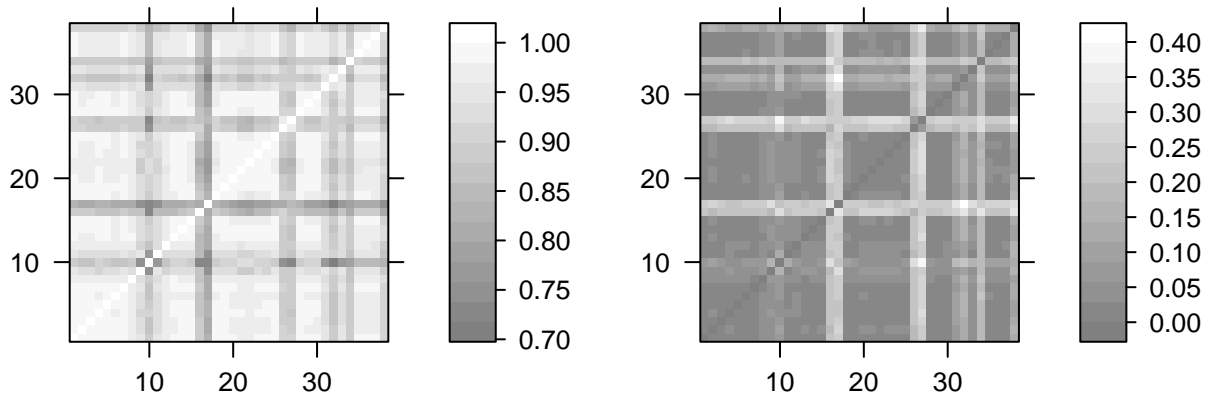


Figure 6: A heatmap of the mean correlation matrix is shown on the left. A heatmap of the difference between the predicted correlation for the maximum speed with the correlation for the minimum observed speed is shown on the right.

The traditional approach to interpreting scores from principal coordinates analysis (or classical multidimensional scaling) is to examine observed values with scores close to those of interest and then try to infer the meaning from these observations. In this case, we can compare the observed correlations closest to minimum and maximum speeds. However, these are single observations and subject to other unrelated kinds of variation. Such a contrast does not reveal the clear meaning shown in the second panel of Figure 6.

6 Discussion

We have shown how data that lie in a non-Euclidean space may be integrated into a regression framework using only a concept of distance and limited assumptions about the space. The ability to backscore from the score space to the data space is essential for the method to be useful for prediction and explanation. The same idea may be expanded to other types of data such images, tensors, trees and data of mixed types where at least a distance may be defined. This is in the spirit of Wang and Marron (2007) who used the term “object-oriented data analysis” and developed some statistical methods for tree data types. The success of such methods will depend on whether a suitable backscoring method can be developed. Circumstances where this is easier involve cases where the data lie on a relatively low dimensional manifold and where it is possible to easily generate weighted averages of pairs of data objects. Even for more familiar continuous or categorical data, we may obtain interesting results with the use of different distance functions.

References

- Cuadras, C. and C. Arenas (1990). A distance based regression model for prediction with mixed data. *Communications in Statistics-Theory and Methods* 19(6), 2261–2279.
- Daniels, M. J. and M. Pourahmadi (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* 89(3), 553–566.
- de Jong, S. (1993). Simpls: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* 18, 251–263.
- Dryden, I. (2009). *shapes: Statistical shape analysis*. R package version 1.1-3.
- Dryden, I. and K. Mardia (1998). *Statistical Shape Analysis*. Chichester: Wiley.
- Dryden, I. L., A. Koloydenko, and D. Zhou (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics* 3(3), 1102–1123.
- Faraway, J. (2012). Backscoring in principal coordinates analysis. *Journal of Computational and Graphical Statistics* 21, 394–412.
- Gower, J. (1968). Adding a point to vector diagrams in multivariate analysis. *Biometrika* 55(3), 582–585.
- Herdin, M., N. Czink, H. Ozcelik, and E. Bonek (2005). Correlation matrix distance, a meaningful measure for evaluation of non-stationary mimo channels. In *Vehicular Technology Conference, 2005. VTC 2005-Spring. 2005 IEEE 61st*, Volume 1, pp. 136 – 140 Vol. 1.
- Higham, N. (2002). Computing the nearest correlation matrix - a problem from finance. *IMA Journal of Numerical Analysis* 22(3), 329–343.
- Legendre, P., F. Lapointe, and P. Casgrain (1994). Modeling brain evolution from behavior: a permutational regression approach. *Evolution* 48, 1487–1499.
- Lichstein, J. W. (2006). Multiple regression on distance matrices: a multivariate spatial analysis tool. *Plant Ecology* 188(2), 117–131.
- McArdle, B. and M. Anderson (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* 82(1), 290–297.
- Mevik, B. and R. Wehrens (2007). The pls package: Principal component and partial least squares regression in R. *Journal of Statistical Software* 18(2), 1–24.
- Ramsay, J. and B. Silverman (2005). *Functional Data Analysis* (2 ed.). New York: Springer.
- Rasmussen, C. and C. Williams (2006). *Gaussian processes for machine learning*. Cambridge, MA: The MIT Press.
- Tenenbaum, J., V. de Silva, and J. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323.
- Trotman, C.-A., J. Faraway, C. Philips, and J. van Aalst (2010). Effects of lip revision surgery in cleft lip/palate patients. *Journal of Dental Research* 89, 728–732.
- Wang, H. and J. Marron (2007). Object Oriented Data Analysis: Sets of Trees. *Annals of Statistics* 35(5), 1849–1873.